AMENDMENTS TO THE CLAIMS:

Claims 1-52 (Cancelled)

53. (Currently Amended) A method for a data processing system to efficiently identify at least one data set of searching for datasets withinfrom a collection of datasets according to a query containing information indicative of desired datasets, the method comprising the machine-executed steps:

constructing a trainable-semantic vector for each dataset;

receiving [[a]] the query containing information indicative of desired datasets;

constructing a trainable-semantic vector for the query;

comparing the trainable-semantic vector for the query to the trainable-semantic vector of each dataset; [[and]]

selecting datasets whose trainable-semantic vectors are closest to the trainable-semantic vector for the query-query; and

generating a result including information of the selected datasets according to a result of the selecting step;

wherein:

the query or each of the datasets includes at least one data point; and

the semantic vector for the query or each of the datasets is constructed by the steps of:

for each data point, constructing a table for storing information indicative of a relationship between each data point and predetermined categories corresponding to dimensions in the semantic space;

2

determining the significance of each data point with respect to the predetermined categories;

constructing a semantic vector for each data point, wherein each semantic vector has dimensions equal to the number of predetermined categories and represents the relative strength of its corresponding data point with respect to each of the predetermined categories; and

combining the semantic vector for each of the at least one data point to form the semantic vector of the query or each of the datasets.

54.   (Original)   The method of Claim 53, wherein the datasets correspond to documents and the query is a natural language query.

55.   (Currently Amended) The method of Claim 53, further comprising the steps:

performing a second search for datasets within the collection of datasets, wherein the second search using a method other than ~~trainable~~ semantic vectors;

combining the two search results to obtain a combined weighted score for each dataset in either of the two search results;

selecting datasets whose combined weighted score is largest.

56.   (Original) The method of Claim 53, further comprising a step of clustering the selected datasets in real time.

57.   (Currently Amended) A method ~~of expanding~~ for efficiently identifying data points in a semantic lexicon related to a dataset, the method comprising the machine-executed steps:

constructing a ~~trainable~~ semantic vector for the dataset;

comparing the ~~trainable~~ semantic vector for the dataset to ~~the trainable~~ semantic ~~vectors~~ vector of each of the data points in [[a]] the semantic lexicon;

3

selecting data points whose ~~trainable~~ semantic vectors are closest to the ~~trainable~~ semantic vector for the dataset; and

adding said selected data points to said ~~dataset.~~ dataset;

wherein:

the dataset includes at least one data point; and

the semantic vector for the dataset is constructed by the steps of:

for each data point, constructing a table for storing information indicative of a relationship between each data point and predetermined categories corresponding to dimensions in the semantic space;

determining the significance of each data point with respect to the predetermined categories;

constructing a semantic vector for each data point, wherein each semantic vector has dimensions equal to the number of predetermined categories and represents the relative strength of its corresponding data point with respect to each of the predetermined categories; and

combining the semantic vector for each of the at least one data point to form the semantic vector of the dataset.

58. (Original) The method of Claim 57, wherein the dataset is a document and the data points are words.

59. (Original) The method of Claim 57, wherein the dataset is a natural language query in a search system and the data points are words.

Claims 60-64 (Cancelled)

4

65.   (Currently Amended) A system for ~~searching datasets within~~identifying at least one data set from a collection of datasets according to a query containing information indicative of desired datasets, the system comprising:

a computer configured to:

construct a ~~trainable~~ semantic vector for each dataset;

receive [[a]] the query containing information indicative of desired datasets;

construct a ~~trainable~~ semantic vector for the query;

compare the ~~trainable~~ semantic vector for the query to the ~~trainable~~ semantic vector of each dataset; [[and]]

select datasets whose ~~trainable~~ semantic vectors are closest to the ~~trainable~~ semantic vector for the ~~query~~ query; and

generate a result including information of the selected datasets according to a result of the selecting step;

wherein:

the query or each of the datasets includes at least one data point; and

the semantic vector for the query or each of the datasets is constructed by the machine-executed steps of:

for each data point, constructing a table for storing information indicative of a relationship between each data point and predetermined categories corresponding to dimensions in the semantic space;

determining the significance of each data point with respect to the predetermined categories;

5

constructing a semantic vector for each data point, wherein each semantic vector has dimensions equal to the number of predetermined categories and represents the relative strength of its corresponding data point with respect to each of the predetermined categories; and

combining the semantic vector for each of the at least one data point to form the semantic vector of the query or each of the datasets.

Claims 66-70 (Cancelled)

71. (Currently Amended) A computer-readable medium carrying one or more sequences of instructions for ~~searching for datasets within~~efficiently identifying at least one data set from a collection of datasets according to an inquiry containing information indicative of desired datasets, wherein execution of the one or more sequences of instructions by one or more processors causes the one or more processors to perform the steps of:

constructing a ~~trainable~~ semantic vector for each dataset;

receiving [[a]] the query containing information indicative of desired datasets;

constructing a ~~trainable~~ semantic vector for the query;

comparing the ~~trainable~~ semantic vector for the query to the ~~trainable~~ semantic vector of each dataset; [[and]]

selecting datasets whose ~~trainable~~ semantic vectors are closest to the ~~trainable~~ semantic vector for the ~~query~~query; and

generating a result including information of the selected datasets according to a result of the selecting step;

wherein:

the query or each of the datasets includes at least one data point; and

6

the semantic vector for the query or each of the datasets is constructed by the steps of:

for each data point, constructing a table for storing information indicative of a relationship between each data point and predetermined categories corresponding to dimensions in the semantic space;

determining the significance of each data point with respect to the predetermined categories;

constructing a semantic vector for each data point, wherein each semantic vector has dimensions equal to the number of predetermined categories and represents the relative strength of its corresponding data point with respect to each of the predetermined categories; and

combining the semantic vector for each of the at least one data point to form the semantic vector of the query or each of the datasets.

Claims 72-75 (Cancelled)